DOCUMENT RESUME

ED 393 882 TM 024 784

AUTHOR Waliczek, Tina M.
TITLE A Primer on Partial Correlation Coefficients.
PUB DATE Jan 96
NOTE 17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Crime; *Demography; Heuristics; *Predictor Variables; *Research Methodology; Urban Areas
IDENTIFIERS *Dependent Variables; Statistical Package for the Social Sciences; *Variance (Statistical)

ABSTRACT

Part and partial correlation coefficients are used to measure the strength of a relationship between a dependent variable and an independent variable while controlling for one or more other variables. The present paper discusses the uses and limitations of partial correlations and presents a small heuristic data set to illustrate the discussion. The correlation of interest was between the number of churches and the number of murders in 15 cities of various populations. The Statistical Package for the Social Sciences was used to analyze the data. Interpretation of the partial correlation value is discussed. It is demonstrated that about 50% of the variance in the number of murders can be associated with the variance in the number of churches when population is being controlled. (Contains one figure, four tables, and six references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 393 882

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
TINA M. WALICZEK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Primer on Partial Correlation Coefficients

Tina M. Waliczek

Texas A&M University 77843-2133

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, January, 1996.

TM024784

BEST COPY AVAILABLE

Abstract

Part and partial correlation coefficients are used to measure the strength of a relationship between a dependent variable and an independent variable while controlling for one or more other variables. The present paper discusses the uses and limitations of partial correlations, and presents a small heuristic data set to illustrate the discussion.

A Primer on Partial Correlation Coefficients

Researchers often look for statistical control in research studies. These scientists are most interested is controlling variance. Statistical control is especially important when a researcher is studying more than one independent variable and the impact these variables have on a dependent variable. Some control can be achieved by using random samples, testing alternative hypotheses, or using a statistical method to isolate the variance in a dependent variable. Part and partial correlations are this type of statistical method. This method allows the researcher to measure the strength of a relationship between a dependent variable and an independent variable while controlling for one or more other variables.

The partial correlation coefficient is a research device employed to examine the linear relationship between three or more variables. Partial correlations are expressed in writing as $r_{xy.z}$. This symbol is interpreted as the correlation between X and Y, while Z is held constant. These correlations are often also expressed as $r_{12.3}$ with the numbers having the same meaning as the letters above.

In many social science research situations, a researcher will be investigating an uncontrolled "real world" environment. Generally, the dependent variable of interest will be predicted on the basis of several independent variable values. The independent variable values, however, are not only related to the dependent variable, but are also related to each other. A partial correlation enables the researcher to examine the relationship among two variables while holding constant the values on one or more other variables.

> In simplest terms, the partial correlation $r_{12.3}$ is merely the correlation between the residual from predicting X_1 from X_3 and the residual from predicting X_2 from X_3. (Hays, 1994, p. 675)

Part correlations are very similar to partial correlations, but the impact of the third variable is "partialed" out from only one of the independent variables. A series of part correlations is often called "stepwise regression" (see Thompson, 1995).

Part and partial correlations are used with interval and ratio data and range between -1 and +1. They are often referred to as "first order", "second order" or "third order" correlations. The term "first", "second", or "third" refers to the number of variables that are being controlled in the correlation. If only one variable is being held constant, then the correlation is called a "first order" correlation. If there are two variables being held constant, then it is referred to as a "second-order" correlation.

A Pearson product-moment correlation is sometimes called a "zero-order correlation", because there are no variables that are being partialed out of the correlation. Part and partial correlations can be computed from the Pearson product-moment correlation coefficients as will be demonstrated in the example problem later in the paper.

Limitations of Partial Correlations

Korn (1984, pp. 61-62) discussed some of the limitations of partial correlations that a researcher should consider. He mentioned mainly that data should be normally distributed and that the relationship between variables should be approximately linear.

> Pearson's partial correlation seems sensitive to the multivariate normality assumption and should probably be used only for

> approximately normally distributed data. However, any partial correlation...will give misleading results when the relationship between the variables is not linear. (Korn, 1984, p. 62)

Partial correlations are only valid when the pattern of relationships between the variables reflects a meaningful model.

> Controlling variables without regard to the theoretical considerations about the pattern of relations among them may amount to a distortion of reality and result in misleading or meaningless results. (Pedhazur, 1982, p. 110)

Two patterns of causation are illustrated in Figure 1. In Figure a, X leads to Y which eventually leads to Z. In Figure b, Y is causing both X and Z. In these types of causal models, it is acceptable to use a partial correlation analysis. Caution should be used when working with variables that have more elaborate patterns of causation.

Another consideration that the researcher may want to be aware of is similar to one involved in the ANCOVA analysis (Benton, 1992, pp. iii-xvii). The researcher must know what dependent variable is being measured after the influence of one or more independent variables are held constant.

Computation of an Example Problem

An example problem is presented here to illustrate the calculation of a partial correlation. The 15 most populated cities in the United States were selected and the populations recorded. Murders that occurred within each city during 1992 were found in the Statistical Abstract of the

United States for 1994. The number of churches and synagogues in each city was also used for the problem. The data are presented in Table 1.

The correlation of interest was between the number of churches and the number of murders in the different cities of various populations. The researcher was additionally interested in finding out if the number of churches and the number of murders in large cities are correlated if the effects of population are controlled. This end was accomplished by calculating a partial correlation. A weighted average between the two variables for each population was attained by computing the partial correlation.

The Statistical Package for Social Sciences (SPSS) was used to analyze the data. A zero-order Pearson product-moment correlation was first completed to analyze each individual correlation presented in Table 2. These correlations showed that all variables are highly correlated with each other. As mentioned earlier, these correlations can be used to compute the partial correlation. The formula for this computation is:

$$r_{XY.Z} = \frac{r_{XY} - (r_{XZ})(r_{YZ})}{\sqrt{1 - r^2_{XZ}}\ \sqrt{1 - r^2_{YZ}}}$$

where, r = correlations between variables

When the Pearson product-moment correlation values in Table 2 were entered into the formula, a partial correlation of .7031 was calculated. This is a moderately high correlation.

$$\frac{.9218 - (.9669)(.8581)}{\sqrt{1-(.9669)^2}\ \sqrt{1-(.8581)^2}}$$

$$\frac{.9218-.8297}{(.2551)(.5135)}$$

$$\frac{.0921}{.1310}$$

$$= .7031$$

This calculation is rather simple, however, and does not illustrate how the partial correlation is obtained by correlating the residuals of variables. A more lengthy calculation is required to demonstrate this idea.

First, it is helpful to obtain the raw scores, predicted scores, and residuals for the 15 cities. The X' and Y' (predicted scores) can be calculated using the regression equation:

$$Y' = a + b\,(z)$$
$$X' = a + b\,(z)$$

where, a = Y- intercept
b = slope
z = population for each city

The y-intercept (a) uses the formula:

$$a = Y - b\,X$$

where, Y = the mean of Y values
X = the mean of X values
b = slope

The slope (b) uses the formula:

$$b = r \frac{SY}{SX}$$

where, r = correlation between variables
SX = standard deviation of X
SY = standard deviation of Y

The means, standard deviations, and correlations were obtained from the computer printouts from the Statistical Package for Social Sciences (SPSS) analysis.

Computations for Y' included:

$$b = r \frac{SY}{SZ}$$

$$b = (.9669) \frac{(515.5)}{(1,753,919)}$$

$$b = (.9669)\ (.00029)$$

$$b = .00028$$

$$a = Y - b\,Z$$

$$a = 467.3 - (.00028)(1,701,963.8)$$

$$a = 467.3 - 476.54$$

$$a = -16.39$$

Computations for X' included:

$$b = r \frac{SX}{SZ}$$

$$b = (.8581) \frac{(890.2)}{(1,753,919)}$$

$$b = (.8581)(.00051)$$

$$b = .00044$$

$$a = X - b\,Z$$

$$a = 1402.3 - (.00044)(1{,}701{,}963.8)$$

$$a = 1402.3 - (748.9)$$

$$a = 660.96$$

Data for b weights and y-intercepts were verified on an SPSS analysis for multiple regression. The final prediction equations were:

$$Y' = -16.39 + (.00028)(Z)$$

$$X' = 660.96 + (.00044)(Z)$$

By filling in each population value into the equations above, the prediction columns of Table 3 were completed. For example, the city of New York's prediction equation for the murder variable (Y) was:

$$Y' = -163.39 + (.00023)(1{,}322{,}564) = 2{,}034 \text{ murders}$$

The predicted value for the number of churches in New York was calculated in this way:

$$X' = 660.96 + (.00044)(7{,}322{,}564) = 3{,}883 \text{ churches}$$

The residual values for the number of churches and the number of murders were calculated once each X' and Y' values were calculated. This was done by taking the actual value for X or Y and subtracting from it the predicted value (X' or Y').

$$Y - Y' = \text{residual} \quad \text{or} \quad X - X' = \text{residual}$$

For example, the actual number of churches and synagogues in New York is 3,505 and the predicted number was 3,883. That leaves a residual value of -378, as reported in Table 3.

$$3505 - 3883 = -378$$

These residualized parts of X and Y are those parts of the values that are not shared with Z. After Z was partialed out, these are the parts of the values that have been left over.

After the residuals were calculated, they were then squared to give a sum of squares value for the predicted X and Y. The residual values for each X and Y variable were then multiplied and summed. These sums were entered into the following equation:

$$r = \frac{(N)(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[(N)(\Sigma X^2) - (\Sigma X)^2][(N)(\Sigma Y^2) - (\Sigma Y)^2}}$$

where, N = number of subjects in data set
Note: when substituting into equation,
X-X' = X and Y - Y' = Y

Using the data from Table 4, the equation can be used to calculate the partial correlation.

$$r = \frac{(15)(590,045) - (-115)(107)}{\sqrt{[(15)(2,927,539) - (-115)^2][(15)(241,169) - (107)^2]}}$$

$$\frac{(8,850,675) - (-12,305)}{\sqrt{[(43,913,085 - 13,225)][(3,617,535 - 11,449)]}}$$

$$\frac{8,862,980}{\sqrt{(43,899,860)(3,606,086)}}$$

$$\frac{8862980}{12{,}581{,}997.88}$$

$$= .7044$$

Although the values are very similar, this partial correlation is not exactly the same as the partial correlation achieved with the use of the Pearson product-moment correlations. This is due to the fact that the values in the tables were rounded to keep calculations as simple as possible. This is, however, a less than 1% difference in the r^2 value of approximately .50.

The partial r^2 value can be interpreted in the same way as a Pearson product-moment correlation. Thus, we can conclude that about 50% of the variance in the number of murders can be associated with the variance in the number of churches in these 15 cities while population is being controlled. Even after the effects of the population variable were controlled, the number of churches have a moderately high correlation with the number of murders in cities.

Literature Cited

Benton, R. (1991). Statistical power considerations in ANOVA. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments. (pp.119-132). Greenwich, CT: AI Press.

Hays, W. L. (1994). Statistics (5th ed.). Austin, TX: Harcourt Brace College Pub.

Korn, E. L. (1984). The ranges of limiting values of some partial correlations under conditional independence. The American Statistician, 38 (1), 61-62.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

United States Department of Commerce (1994). Statistical abstract of the United States 1994 (114th ed.).

Table 1: Raw data on 15 cities

ID	City	Religion (X)	# Murders (Y)	Population (Z)
1	New York	3,505	1,984	7,322,564
2	Los Angeles	2,023	1,056	3,485,557
3	Chicago	2,863	921	2,783,726
4	Houston	2,011	447	1,629,902
5	Philadelphia	1,709	420	1,585,577
6	San Diego	582	141	1,110,623
7	Detroit	1,475	586	1,027,974
8	Dallas	1,313	373	1,007,618
9	Phoenix	663	134	983,403
10	San Antonio	937	211	935,393
11	San Jose	355	42	782,224
12	Indianapolis	1,010	132	741,952
13	Baltimore	1,098	326	736,014
14	San Francisco	615	113	723,959
15	Jacksonville	875	125	672,971

Table 2: Correlation Coefficients

	# Murders (Y)	Religion (X)	Population (Z)
# Murders (Y)	1.0000 (15) P=.	.9218 (15) P=.000	.9669 (15) P=.000
Religion (X)	.9218 (15) P=.000	1.0000 (15) P=.	.8581 (15) P=.000
Population (Z)	.9669 (15) P=.000	.8581 (15) P=.000	1.0000 (15) P=.

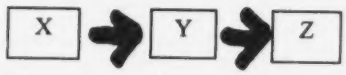
Table 3: Raw scores, predicted scores, and residuals for 15 cities

ID	City	Religion (X)	# Murders (Y)	Population (Z)	X' with Z independent	Y' with Z independent	X-X'	Y-Y'
1	New York	3,505	1,984	7,322,564	3,883	2,034	-378	-50
2	Los Angeles	2,023	1,056	3,485,557	2,195	960	-172	96
3	Chicago	2,863	921	2,783,726	1,886	763	977	158
4	Houston	2,011	447	1,629,902	1,378	440	633	7
5	Philadelphia	1,709	420	1,585,577	1,359	428	350	-8
6	San Diego	582	141	1,110,623	1,150	295	-568	-154
7	Detroit	1,475	586	1,027,974	1,113	271	362	315
8	Dallas	1,313	373	1,007,618	1,104	266	209	107
9	Phoenix	663	134	983,403	1,094	259	-431	-125
10	San Antonio	937	211	935,393	1,073	246	-136	-35
11	San Jose	355	42	782,224	1,005	203	-650	-161
12	Indianapolis	1,010	132	741,952	987	191	23	-59
13	Baltimore	1,098	326	736,014	985	190	113	136
14	San Francisco	615	113	723,959	980	186	-365	-73
15	Jacksonville	875	125	672,971	957	172	-82	-47

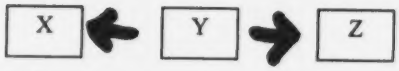
Table 4: Computational procedures for the partial correlation coefficient using residuals

City ID	Regression residual for religion with population independent X-X'	Regression residual for number of murders with population independent Y-Y'	$(X-X')^2$	$(Y-Y')^2$	$(X-X')(Y-Y')$
1	-378	-50	142,884	2,500	18,900
2	-172	96	29,584	9,216	-16,512
3	977	158	954,529	24,964	154,366
4	633	7	400,689	49	4,431
5	350	-8	122,500	64	-2,800
6	-568	-154	322,624	23,716	87,472
7	362	315	131,044	99,225	114,030
8	209	107	43,681	11,449	22,363
9	-431	-125	185,761	15,625	53,875
10	-136	-35	18,496	1,225	4,760
11	-650	-161	422,500	25,921	104,650
12	23	-59	529	3,481	-1,357
13	113	136	12,769	18,496	15,368
14	-365	-73	133,225	5,329	26,645
15	-82	-47	6,724	2,209	3,854
	Σ= -115	Σ= 107	Σ= 2,927,539	Σ= 241,169	Σ= 590,045

Figure 1:



(a)



(b)